Peer Review File

Manuscript Title: Discovery of bat coronaviruses close to SARS-CoV-2 and infectious for human cells

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referee #1:

This paper is a tour de force. It establishes that Rhinolophus bats harbour SARS-CoV-2-related viruses with a receptor binding domain (RBD) almost identical to the human virus, with comparably high affinity for human ACE2. I have just a few suggestions for improvement:

1. It would be good to state clearly that BANAL-52 is the most similar animal virus to SARS-CoV-2 and to give its similarity (96.8% by my calculation) compared to RaTG13 (96.1).

2. It would be good to add a phylogenetic tree based just on the RBD to highlight how much closer BANAL viruses are to SARS-CoV-2 in that domain compared to RaTG13.

3. There should be a clear statement in the main text of how many Rhinolophus bats were sampled versus how many sarbecoviruses were identified.

It was an honour and a pleasure to review this paper!

-Michael Worobey

Referee #3:

Temmam and colleagues report the identification of a SARS-CoV-2-like virus in bats that harbors a receptor binding domain (RBD) very similar to that of SARS-CoV-2. Specifically, they show that pseudotypes bearing the spike protein of this bat virus are able to efficiently enter cells transfected to express human ACE2. Furthermore, isolation of the authentic virus on VeroE6 cells, which express endogenous ACE2, is shown. These findings are of high interest to the field. However, some points remain to be addressed.

Major

293T cells transfected with an ACE2 expression plasmid express very high amounts of ACE2 that will allow robust entry even of particles bearing spike proteins from sarbecoviruses that use human ACE2 with low efficiency. Therefore, it is essential to determine whether the BANAL-236 spike also mediates entry into human cell lines (preferentially lung derived) that express endogenous ACE2.

One might argue that isolation of BANAL-236 on VeroE6 cells, which express endogenous ACE2, is sufficient proof for BANAL-236 using endogenous ACE2. However, VeroE6 cells are of African green monkey and not human origin and express high amounts of ACE2. Furthermore, no attempt was made to confirm whether propagation of BANAL-236 in VeroE6 cells was ACE2 dependent. Therefore, only limited conclusions can be drawn from the data. It would be highly desirable to test whether authentic BANAL-236 infects human cell lines known to express endogenous ACE2 and known to be susceptible to SARS-CoV-2 infection.

Minor

Figure 4B shows that particles bearing BANAL-236 spike like their counterparts bearing SARS-CoV-2 spike are neutralized by sera from COVID-19 patients but not control sera. Formally, a negative control – particles bearing an unrelated viral glycoprotein like VSV-G – should be included to demonstrate specificity of neutralization. Furthermore, titration of sera is desirable in order to reveal potential differences in neutralization of particles bearing BANAL-236 and SARS-CoV-2 spike.

Please indicate whether a single representative experiment or the average of several independent experiments is shown. If a single experiment is shown, please state the number of confirmatory experiments. If the average is shown, please state how many experiments were averaged. Finally, please indicate whether error bars indicate SEM or SD.

Referee #4:

The manuscript entitled " Coronaviruses with a SARS-CoV-2-like receptor-binding domain allowing ACE2-mediated entry into human cells isolated from bats of Indochinese peninsula" described the discovery of SARS-CoV-2-like coronaviruses from bats in south Asia that share a very similar receptor-binding domain with SARS-CoV-2. The finding and the writing are both straightforward. The study makes an important contribution to clarifying the origin of COVID-19. I don't have any major issues with the manuscript. Here are some minor comments:

(1) The title is difficult to parse and should be rephrased. There are also some other awkward sentences in the manuscript that can be improved.

(2) lines 68-69: "no SARS-CoV-2-like virus has been shown to use hACE2 to enter human cells". This is not true. RaTG13 spike protein has been shown to use hACE2 for human cell entry (although inefficiently). Also, would pangolin coronaviruses be considered SARS-CoV-2-like viruses? Consider changing the sentence to "no SARS-CoV-2-like virus from bats has been shown to use hACE2 to enter human cells efficiently".

(3) lines 250-252: "RaTG13 RBD, the virus identified in R. affinis from the Mojiang mineshaft where pneumonia cases with clinical characteristics strikingly similar to COVID-19 were recorded in 2012." This sentence has a high possibility of been misinterpreted. Consider removing "where pneumonia cases with clinical characteristics strikingly similar to COVID-19 were recorded in 2012", since this is an opinion rather than a fact.

(4) Section on the recombination events: are these recombination sites random or do they contain some signature motifs? Why the beginning of the RBD is a breakpoint? Can the authors include a comment on this?

(5) Fonts in the figures (particularly Figure 1-3) are too small. Once in print, readers will have a hard time reading the figures. The authors need to find ways to enlarge the fonts.

(6) Figure 3A: Instead of citing the binding affinity between SARS-CoV-2 RBD and hACE2, why not include SARS-CoV-2 RBD in the biolayer interferometry assay? It is hard to compare the binding data from different studies because the experimental conditions may differ among the studies.

(7) Figure 3C: If there is no amino acid difference in helix H4 between BANAL and SARS-CoV-2 RBDs, why would this helix have different structures in the two RBDs? Could the structural difference result from modelings? Unbiased electron density (e.g., composite omit map) of helix H4 needs be included in the supplementary data. Moreover, unbiased electron densities of the critical interacting residues at the interface (e.g., H498 and K493 and residues interacting with them) also need to be shown.

(8) The Q498H mutation has been studied extensively in the literature. Please cite some literature

on this mutation.

(9) Figure 4A: BABAL-236 pseudovirus seems to use hACE2 more efficiently than SARS-CoV-2 pseudovirus. This result may be misleading because the pseudoviruses were not calibrated for the expression levels of the spike proteins. A Western blot on the expressed spike proteins in the pseudoviruses needs to be shown.

Referee #5:

This study provides new insights into the evolutionary history of SARS-CoV-2. By sampling from 645 bats from 6 families in northern Laos, the authors discover 5 Sarbecoviruses in Rhinolophus bats that are related to SARS-CoV-2. The Laos viruses are not the most genetically similar to SARS-CoV-2 at a whole-genome level, but they share key residues in the receptor-binding domain that allow the bat virus to bind to human cellular receptors (hACE2). This is one of the most important evolutionary adaptations required for spillover into humans and this study shows that the motif occurs naturally in bats; thus an intermediary host is not required.

Technical comments:
1. Figure S4: There seems to be a problem in the alignment around the S1/S2 cleavage site. The NSPA/NSPV amino acids should be aligned with NSPR near the cleavage site, not with QTQT as they appear in the figure. When the alignment is corrected, the BANAL viruses 20-116 and 20-247 do have an insertion at the S1/S2 cleavage site, similar to the RmYN02 bat virus described in Zhou et al. Current Biology.
2. When you selected 36 of the original 106 Sarbecovirus sequences for the recombination analysis, how do you know that you did not incidentally exclude recombinant viruses that may have had a region similar to SARS-CoV-2, even if most the genome was dissimilar? Shouldn't the recombination analysis be performed prior to the downsampling of the data?
3. Line 117: How do these breakpoints compare to those defined in recent studies of Sarbecovirus evolution? How many are new?
4. There is scant information on the methods for the phylogenetic analysis. What model of nucleotide substitution was used? Why weren't bootstrap values inferred for such a small dataset.
5. Can you describe the coverage you achieved in sequencing, particularly for key sites of interest?
6. Lines 95-97: it seems like a waste to throw out data from viruses with 90% genome coverage, given how few data points we have. Was there good coverage of the spike protein?
7. Line 189-190: how conserved were SARS-CoV-2 epitopes in the BANAL viruses?

Manuscript organization:
1. A summary table is needed in the main text that shows how the BANAL viruses share important residues in the RBD that contact with ACE2, compared to other bat and pangolin viruses, rather than burying this key data in the supplement (show data for cleavage site as well).
2. It would also be helpful to have some simple summary table or chart of genetic similarity between Laos viruses, SARS-CoV-2, RaTG13 and select other bat viruses, including (a) whole-genome, (b) spike, (c) RBD, etc.
3. The Discussion section is unnecessarily long and at times hyperbolic and needs to be tightened and focused. For example, this study provides some important insights but it is inaccurate to suggest in line 221 that these data show that RaTG13 is no longer the "proximal ancestor" of SARS-CoV-2. For quite some time it's been shown that RaTG13 only contributed sections of the SARS-CoV-2 genome and other viruses like RmYN02 are most closely related in other sections. Also rephrase line 247. The BANAL viruses are similar to SARS-CoV-2 in an important way but they are not "very close".

Minor clarifications:
1. It would be helpful in the supplement to have maps of the geographic ranges of these Rhinolophus bat species (mashalli, malaynus, and pusillus)
2. There is a typo in the Figure S4 legend: "clivage site" should be "cleavage site"
3. The resolution in Figure 1D is poor. There should be more resolution for the higher part of the

scale. Otherwise all that is being highlighted is the lower AA identity data points. which are not of interest.
4. This line in the Abstract is not accurate: "Several studies have suggested the involvement of pangolin coronaviruses in SARS-CoV-2 emergence7–9 Better to simply say that viruses with similar genomic features as SARS-CoV-2 have been identified in pangolins.
5. Did you also test the bat saliva or urine samples?
6. Line 57: "evolved only through sporadic mutations" not true, recombination (Jackson et al 2021 Cell)
7. GenBank accession numbers are listed for the 5 BANAL viruses that are the focus of this study. Did you sequence and deposit the genomes of the other 19 coronaviruses from bats? These are valuable data. The next bat coronavirus that infects humans could come from any part of the tree and it's helpful to have as many background sequences as possible as context.

Referee #6:

The work reported in this manuscript sought to identify bat viruses similar to SARS-CoV-2 in order to further probe the origin of the virus that led to the COVID-19 pandemic. I have been asked to provide peer-review on its computational component, which focused on three such identified viruses (BANAL 52, 103, 236).

Based on the X-ray structure 6M0J, which describes the SARS-CoV-2 spike RBD complexed with a fragment of the ACE2 receptor, homology models were constructed for BANAL 52/103 and 236, and atomistic simulations were performed in triplicate for the models and 6M0J. Given sequence identity ~97% within the modeled regions, homology modeling is highly appropriate. Reported MODELLER and GROMACS protocols are reasonable, and sampling time reaches the microsecond timescale.

It was unclear to me why the authors retained crystallographic ions in the models, but not waters, which can be important for mediating interactions in complexes. It was unclear why they did not perform their simulations at physiological temperature. A more significant concern is that all of the models describe fragments, which are at high risk of unfolding in the absence of restraints, especially on long timescales, especially for a system that contains unstructured regions like the RBD; yet, the authors do not indicate the use of restraints during their production runs. They do describe the use of distance restraints "to ensure the formation of 7 disulfide bonds" in the template, and this is very odd. A disulfide bond should be an actual S-S bond, with specific force field parameters to define its bond-angle-dihedral behavior; disulfide bonds are indeed available in CHARMM36m. While these distance-based restraints likely helped to prevent unfolding of the RBD fragment, the disulfide dynamics are modeled incorrectly if treated as restraints, and with so many for so small a fragment this could influence the dynamics of the entire RBD-ACE2 model.

The reader could be assured that the simulations are stable and the production runs converged if plots of RMSD over time were presented, particularly for the entire complex. The current RMSD plots are presented as distributions, and not all of the distributions are visible in the plots presented due to overlap. One cannot know from this presentation if the RMSD is converging, diverging, or simply fluctuating over the course of the simulation. Still it is clear from the irregular and broad RMSD distributions that the BANAL 52/103 interface has quite different behavior from the other models. In the absence of further information, this reader questions the stability of the interface. The different behavior of BANAL 52/103 leads to the identification of two other clusters that describe the interface, and the reason for this is not discussed–this is a key point, if the one model exhibits multiple clusters, its behavior is not in fact "identical" to the others, as the authors claim. Another way to communicate simulation findings to readers would be through the presentation of a supplemental movie of the simulation, which would be welcome here. A number of other MD simulations have already been published from 6M0J; perhaps the authors could support their case also by citing others.

The authors use ROSETTA and FoldX scores to approximate binding energies and report these data

as distributions, which are again difficult to decipher because of overlap. For what is visible, there appears to be a range of values across which the distributions are centered. An alternative presentation would allow readers to judge for themselves how similar the data really are.

After carrying out all of this computational work, the authors make very little comment on it, note no caveats or limitations, and do not refer to it during discussion. While the computational aspect of this work is not novel, and does not contribute significantly to the conclusions of the study, it appears sufficiently reliable for its purpose. While the actual interaction between the SARS-CoV-2 spike and ACE2 receptor depends on many other factors that are not considered in the simplified computational models used here, the work does indicate that the isolated binding interfaces of SARS-CoV-2 and BANAL 52/103 and 236 share some notable similarities over one-microsecond timescales.

Author Rebuttals to Initial Comments:

Referee #1 (Remarks to the Author):

This paper is a tour de force. It establishes that Rhinolophus bats harbour SARS-CoV-2-related viruses with a receptor binding domain (RBD) almost identical to the human virus, with comparably high affinity for human ACE2. I have just a few suggestions for improvement:

1. It would be good to state clearly that BANAL-52 is the most similar animal virus to SARS-CoV-2 and to give its similarity (96.8% by my calculation) compared to RaTG13 (96.1).

Done, the statement was added to the main text, line 113.

2. It would be good to add a phylogenetic tree based just on the RBD to highlight how much closer BANAL viruses are to SARS-CoV-2 in that domain compared to RaTG13.

We thank the reviewer for this suggestion. The tree based on RBD amino-acid sequences was placed in place of the tree based on whole genome nucleotide sequences in Fig.1 which is less informative because of the mosaicism of the genome. We moved the whole genome phylogenetic tree in Supp. Figure 2, before the different recombinant fragments.

3. There should be a clear statement in the main text of how many Rhinolophus bats were sampled versus how many sarbecoviruses were identified.

This information is provided in the main section "Diversity of bats and coronaviruses in caves" and in Table S1.

It was an honour and a pleasure to review this paper!

-Michael Worobey

Referee #2 (Remarks to the Author):

Temmam and colleagues report the identification of a SARS-CoV-2-like virus in bats that harbors a receptor binding domain (RBD) very similar to that of SARS-CoV-2. Specifically, they show that pseudotypes bearing the spike protein of this bat virus are able to efficiently enter cells transfected to express human ACE2. Furthermore, isolation of the authentic virus on VeroE6 cells, which express endogenous ACE2, is shown. These findings are of high interest to the field. However, some points remain to be addressed.

Major

293T cells transfected with an ACE2 expression plasmid express very high amounts of ACE2 that will allow robust entry even of particles bearing spike proteins from sarbecoviruses that use human ACE2 with low efficiency. Therefore, it is essential to determine whether the BANAL-236 spike also mediates entry into human cell lines (preferentially lung derived) that express endogenous ACE2.

One might argue that isolation of BANAL-236 on VeroE6 cells, which express endogenous ACE2, is sufficient proof for BANAL-236 using endogenous ACE2. However, VeroE6 cells are of African green monkey and not human origin and express high amounts of ACE2. Furthermore, no attempt was made to confirm whether propagation of BANAL-236 in VeroE6 cells was ACE2 dependent. Therefore, only limited conclusions can be drawn from the data. It would be highly desirable to test whether authentic BANAL-236 infects human cell lines known to express endogenous ACE2 and known to be susceptible to SARS-CoV-2 infection.

In the revised version of the manuscript, we now show that authentic BANAL-236 infects human cell lines that express endogenous ACE2 such as Calu-3 (lung cells) and Caco-2 (intestinal cells) (Fig. 4C). Cells were infected at a MOI of 0.01 in the absence of TPCK and genome copy number was quantified by RT-qPCR in the supernatants recovered 3- and 4-days post-infection. BANAL-236 replicates in these cells albeit with a slower kinetic compared to Wuhan. BANAL-236 viral titers obtained at day 4 are reported.

In order to confirm that the propagation of BANAL-236 in VeroE6 cells was ACE2 dependent, we preincubated the viral inoculum with soluble ACE2 (sACE2) at 25 µg/mL (from doi.org/10.1016/j.cell.2020.04.004) for 30 min and infected VeroE6 cells at a MOI of 0.0001 (Fig. 4C). Genome copy number of BANAL-236 was decreased by 3.6 log in the presence of sACE2 at day 3, thus showing the ACE2-dependent entry of this virus.

Minor

Figure 4B shows that particles bearing BANAL-236 spike like their counterparts bearing SARS-CoV-2 spike are neutralized by sera from COVID-19 patients but not control sera. Formally, a negative control – particles bearing an unrelated viral glycoprotein like VSV-G – should be included to demonstrate specificity of neutralization. Furthermore, titration of sera is desirable in order to reveal potential differences in neutralization of particles bearing BANAL-236 and SARS-CoV-2 spike.

We did not perform serum titration because these sera are valuable and we plan to use them in other set of experiments using neutralization of pseudotypes, plaque reduction and binding tests against different domains of the different spikes. We think that quantifying cross-neutralization as suggested is out of the scope of the current paper.

Please indicate whether a single representative experiment or the average of several independent experiments is shown. If a single experiment is shown, please state the number of confirmatory experiments. If the average is shown, please state how many experiments were averaged. Finally, please indicate whether error bars indicate SEM or SD.

The following information has been added to the legend of Fig. 4:

- *Figure 4A: A single experiment performed in triplicate representative of 2 experiments is shown. Error bars indicate SD.*
- *Figure 4B: A single experiment representative of 3 independent experiments is shown.*

Referee #3 (Remarks to the Author):

The manuscript entitled "Coronaviruses with a SARS-CoV-2-like receptor-binding domain allowing ACE2-mediated entry into human cells isolated from bats of Indochinese peninsula" described the discovery of SARS-CoV-2-like coronaviruses from bats in south Asia that share a very similar receptor-binding domain with SARS-CoV-2. The finding and the writing are both straightforward. The study makes an important contribution to clarifying the origin of COVID-19. I don't have any major issues with the manuscript. Here are some minor comments:

(1) The title is difficult to parse and should be rephrased. There are also some other awkward sentences in the manuscript that can be improved.

The title has been rephrased: "Discovery of bat coronaviruses in Indochinese peninsula highly similar to SARS-CoV-2 sheds light onto the origin of the COVID-19 pandemic".

(2) lines 68-69: "no SARS-CoV-2-like virus has been shown to use hACE2 to enter human cells". This is not true. RaTG13 spike protein has been shown to use hACE2 for human cell entry (although inefficiently). Also, would pangolin coronaviruses be considered SARS-CoV-2-like viruses? Consider changing the sentence to "no SARS-CoV-2-like virus from bats has been shown to use hACE2 to enter human cells efficiently".

*We modified the sentence accordingly: "no **bat** SARS-CoV-2-like virus has been shown to use hACE2 to **efficiently** enter human cells" (lines 69-70).*

(3) lines 250-252: "RaTG13 RBD, the virus identified in R. affinis from the Mojiang mineshaft where pneumonia cases with clinical characteristics strikingly similar to COVID-19 were recorded in 2012." This sentence has a high possibility of been misinterpreted. Consider removing "where pneumonia cases with clinical characteristics strikingly similar to COVID-19 were recorded in 2012", since this is an opinion rather than a fact.

*Rather than deleting, the sentence was rephrased, and we have softened the assertion as follows: "the RaTG13 RBD, the virus identified in R. affinis from the Mojiang mineshaft "where pneumonia cases with clinical characteristics **a posteriori interpreted as similar to COVID-19** were recorded in 2012" (lines 257-258).*

(4) Section on the recombination events: are these recombination sites random or do they contain some signature motifs? Why the beginning of the RBD is a breakpoint? Can the authors include a comment on this?

The RBD is located seven nucleotides from the breakpoint n°10 identified by GARD (fragment 11 comprised the region 22389-24229 in the alignment while the RBD comprised the region 22396-23037). This feature was added to the main text (lines 131-132).

No specific signature was identified in the different breakpoints. Of note, all breakpoints were located in coding regions and most of them (9/14) targeting the second position of the codon (see the table below presenting the principal features of the breakpoints; this table was added to the Supplementary data [Table S7]):

Breakpoint no	***Downstream fragment no***	***Coordinates of the fragment***	***Breakpoint located in coding region?***	***Breakpoint position in the codon***	***Sequences before (aa)***	***Breakpoint sequence***	***Sequences after (aa)***
1	*2*	*1 – 1,820*	*yes (ORF1a)*	*3rd*	*GGTGGTGTTGTTCA (GGVVQ)*	*G*	*TTGACTTCGCAGTGG (LTSQW)*
2	*3*	*1,821 – 2,689*	*yes (ORF1a)*	*2nd*	*CCACTGGGCATTG (PLGID)*	*A*	*TTAGATGAGTGGAGT (LDEWS)*
3	*4*	*3,971 – 7,176*	*yes (ORF1a)*	*2nd*	*GTGGTTATACCTA (VVIPT)*	*C*	*TAAAAAGGCTGGTGGC (KKAGG)*
4	*5*	*7,177 – 11,461*	*yes (ORF1a)*	*1st*	*ATGTGGTTAATAATT (MWLIII)*	*A*	*ATCTTGTACAAATG (NLVQM)*
5	*6*	*11,462 – 12562*	*yes (ORF1a)*	*2nd*	*TTTGGCCTCTTTT (FGLFC)*	*G*	*TTTACTCAACCGCTAC (LLNRY)*
6	*7*	*12,563 – 17694*	*yes (ORF1a)*	*2nd*	*ACTGCTTGCACTG (TACTD)*	*A*	*TGACAATGCGTTAGCT (DNALA)*
7	*8*	*17,695 – 19,791*	*yes (ORF1ab)*	*2nd*	*ACCACTGAAACAG (TTETA)*	*C*	*AGCTCACTCTTGTAAT (AHSCN)*
8	*9*	*19,792 – 21,204*	*yes (ORF1ab)*	*2nd*	*CTCACTGTCTTTT (LTVFF)*	*T*	*TGATGGTAGAGTTGAT (DGRVD)*
9	*10*	*21,205 – 22,388*	*yes (ORF1ab)*	*2nd*	*TTATTTGACATGA (LFDMS)*	*G*	*TAAATTTCCCCTTAAA (KFPLK)*
10	*11*	*22,389 – 24,229*	*yes (S)*	*3rd*	*ACAGAATCTATTGT (TESIV)*	*T*	*AGATTTCCTAATATT (RFPNI)*
11	*12*	*24,230 – 27,343*	*yes (S)*	*2nd*	*CAAGACTCACTTT (QDSLS)*	*C*	*TTCCACAGCAAGTGCA (STASA)*
12	*13*	*27,344 – 27,800*	*yes (ORF7a)*	*2nd*	*GAGTGTGTTAGAG (ECVRG)*	*G*	*TACAACAGTACTTTTA (TTVLL)*
13	*14*	*27,801 – 28,463*	*yes (ORF8)*	*1st*	*CTTGTTTTCTTAGGA (LVFLG)*	*A*	*ATCATCACAACTGTA (IITTV)*
14	*15*	*28,464 – 29,597*	*yes (N)*	*3rd*	*AGAGCTACCAGACG (RATRR)*	*A*	*ATTCGTGGTGGTGAC (IRGGD)*

(5) Fonts in the figures (particularly Figure 1-3) are too small. Once in print, readers will have a hard time reading the figures. The authors need to find ways to enlarge the fonts.

Figures 1- 3 were reformatted to take into account reviewer's comment: especially the bottom part of figure 2 presenting a subset of phylogenies performed on recombinant fragments was deleted since the complete phylogenies are presented in Supp. Figure 2. In addition, all the phylogenies performed on the 15 recombinant fragments were updated to follow the color code of Figure 2.

(6) Figure 3A: Instead of citing the binding affinity between SARS-CoV-2 RBD and hACE2, why not include SARS-CoV-2 RBD in the biolayer interferometry assay? It is hard to compare the binding data from different studies because the experimental conditions may differ among the studies.

We thank the reviewer for raising this point. We have now carried out biolayer interferometry experiments in parallel, directly comparing the binding of BANAL and SARS-CoV-2 RBDs to hACE2 and found that RBDs 52/103 and 236 have a binding affinity for hACE2 that is three times higher than that of SARS-CoV-2 RBD. This higher affinity can be attributed exclusively to the Q498H difference found in both BANAL RBDs. The structure shows that H498 is at 3.5Å distance of D38, probably forming a hydrogen bond. An affinity increase by a mutation H498Q in the spike of SARS-CoV-2 has indeed been reported in previous studies inducing a more positively charged RBM surface that can form a stronger interaction with the negatively charged surface of hACE2. The Q493K difference present in RBD 236 seems not to have a strong influence the K_D of RBD 236 for human ACE2 is similar to that of RBD 52/103. The main text and the new Figure 3 have been revised accordingly.

(7) Figure 3C: If there is no amino acid difference in helix H4 between BANAL and SARS-CoV-2 RBDs, why would this helix have different structures in the two RBDs? Could the structural difference result from modelings? Unbiased electron density (e.g., composite omit map) of helix H4 needs be included in the supplementary data. Moreover, unbiased electron densities of the critical interacting residues at the interface (e.g., H498 and K493 and residues interacting with them) also need to be shown.

Following this request, we now present a composite omit map for the segment D364-S375 of RBD 236, which confirms the correct assignment of this region. Concerning the reviewer's question, we note that the change A372T in BANAL RBDs converts the ^{370}NSA^{372} motif into ^{370}NST^{372}, an N-linked-glycosylation sequon (Ref 1-3). The electron density map of the RBD236-hACE2 complex indeed had extra density extending from N370, which we assigned to the first N-acetylglucosamine residue of an attached glycan. Our refinement shows that the main chain carbonyl group of T372 makes a hydrogen bond with ring Oxygen atom linking carbons 1 and 5 of the NAG residue (marked in the Figure). This altered orientation of the main chain at T372 propagates downstream, inducing a different conformation downstream of helix 4.

(8) The Q498H mutation has been studied extensively in the literature. Please cite some literature on this mutation.

We now cite the various reports mentioning this mutation, which are also listed here for the reviewers' attention:

20. *Niu S, Wang J, Bai B, Wu L, Zheng A, Chen Q, Du P, Han P, Zhang Y, Jia Y, Qiao C, Qi J, Tian WX, Wang HW, Wang Q, Gao GF. Molecular basis of cross-species ACE2 interactions with SARS-CoV-2-like viruses of pangolin origin. EMBO J. 2021 Aug 16;40(16):e107786.*

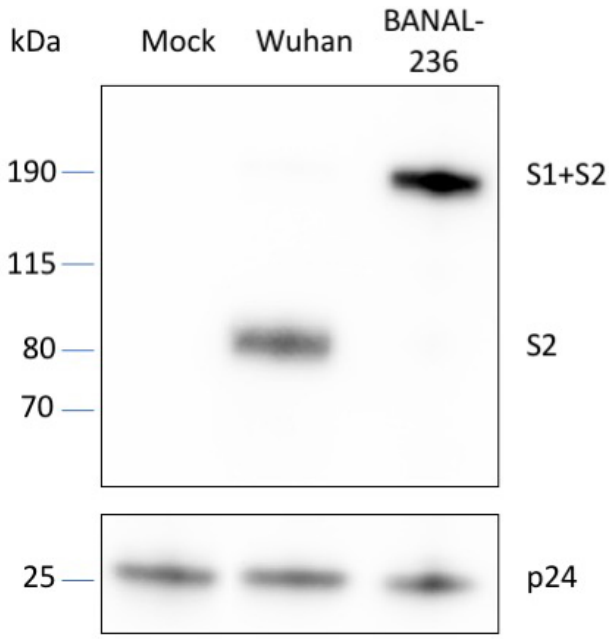
21. *Zhang Y, Huang K, Wang T, Deng F, Gong W, Hui X, Zhao Y, He X, Li C, Zhang Q, Chen X, Lv C, Lin X, Yang Y, Sun X, Shi Z, Chen H, Zou Z, Jin M. SARS-CoV-2 rapidly adapts in aged BALB/c mice and induces typical pneumonia. J Virol. 2021 Mar 10;95(11):e02477-20.*

22. *Huang K, Zhang Y, Hui X, Zhao Y, Gong W, Wang T, Zhang S, Yang Y, Deng F, Zhang Q, Chen X, Yang Y, Sun X, Chen H, Tao YJ, Zou Z, Jin M. Q493K and Q498H substitutions in Spike promote adaptation of SARS-CoV-2 in mice. EBioMedicine. 2021 May;67:103381.*

23. *Zhang S, Qiao S, Yu J, Zeng J, Shan S, Tian L, Lan J, Zhang L, Wang X. Bat and pangolin coronavirus spike glycoprotein structures provide insights into SARS-CoV-2 evolution. Nat Commun. 2021 Mar 11;12(1):1607.*

(9) Figure 4A: BANAL-236 pseudovirus seems to use hACE2 more efficiently than SARS-CoV-2 pseudovirus. This result may be misleading because the pseudoviruses were not calibrated for the expression levels of the spike proteins. A Western blot on the expressed spike proteins in the pseudoviruses needs to be shown.

In the revised version of the manuscript, we provide a western-blot analysis of the spike proteins expressed in the pseudotyped lentiviral particles (Supp Figure 10). For an equivalent amount of p24, lentiviral particles pseudotyped with S-BANAL-236 expressed a higher level of spike protein compared to lentiviral particles pseudotyped with S-Wuhan, which could explain the higher transduction efficiency of BANAL-236 pseudoviruses on 293T-ACE2 in Fig. 4A. We did not claim in the manuscript a higher efficiency in cell entry of BANAL-236 and, due to restricted space, we did not comment on relative efficiency.



Referee #4 (Remarks to the Author):

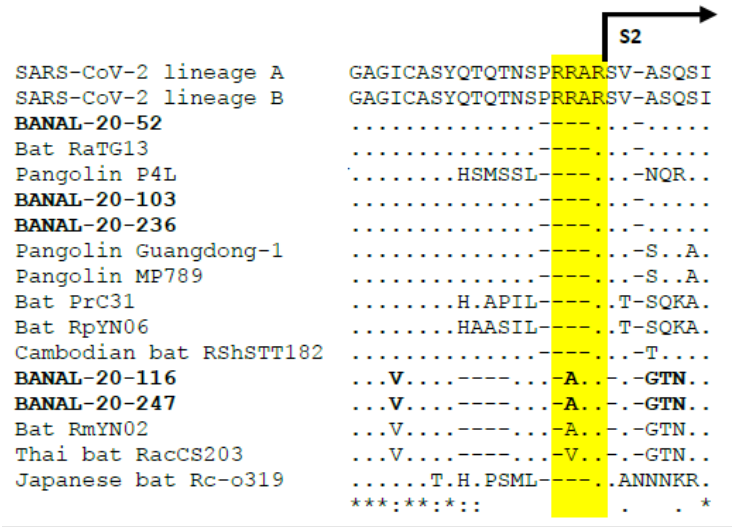
This study provides new insights into the evolutionary history of SARS-CoV-2. By sampling from 645 bats from 6 families in northern Laos, the authors discover 5 Sarbecoviruses in Rhinolophus bats that are related to SARS-CoV-2. The Laos viruses are not the most genetically similar to SARS-CoV-2 at a whole-genome level, but they share key residues in the receptor-binding domain that allow the bat virus to bind to human cellular receptors (hACE2). This is one of the most important evolutionary adaptations required for spillover into humans and this study shows that the motif occurs naturally in bats; thus an intermediary host is not required.

The authors want to highlight that at the whole genome level, BANAL-52 (96.8% nucleotide identity) is closer to SARS-CoV-2 Wuhan-Hu-1 prototype strain than RaTG13 (96.1%). We agree that this information was not clearly stated in the manuscript, and added the information accordingly (lines 107-108).

Technical comments:

1. Figure S4: There seems to be a problem in the alignment around the S1/S2 cleavage site. The NSPA/NSPV amino acids should be aligned with NSPR near the cleavage site, not with QTQT as they appear in the figure. When the alignment is corrected, the BANAL viruses 20-116 and 20-247 do have an insertion at the S1/S2 cleavage site, similar to the RmYN02 bat virus described in Zhou et al. Current Biology.

The most parsimonious alignment generated by MAFFT (parameter G-INS-i), meaning the one that minimized the gap open penalty, gave the result presented in Figure S4. However, we agree that BANAL-116 and BANAL-247 are phylogenetically close to RmYN02, and below the reviewer will find the alignment of the amino-acid furin cleavage region as suggested by the reviewer. At the nucleotide level, Spyros Lytras presented another version of the alignment of this key region (see https://virological.org/t/the-sarbecovirus-origin-of-sars-cov-2-s-furin-cleavage-site/536/5) that conserved part of the furin cleavage site. Since we cannot determine which alignment is correct, we



2. When you selected 36 of the original 106 Sarbecovirus sequences for the recombination analysis, how do you know that you did not incidentally exclude recombinant viruses that may have had a region similar to SARS-CoV-2, even if most the genome was dissimilar? Shouldn't the recombination analysis be performed prior to the downsampling of the data?

The selection of the 36 representative sequences was based on a preliminary recombination analysis performed with 106 complete sarbecovirus genomes with RDP4.0 software and confirmed by phylogenetic analyses. This analysis identified several breakpoints and subsequent phylogenetic analyses revealed that none of the SARS-CoV-1-related viruses was contributive to SARS-CoV-2; therefore, these sequences were removed from the final analysis.

3. Line 117: How do these breakpoints compare to those defined in recent studies of Sarbecovirus evolution? How many are new?

The identification of the breakpoints is highly dependent on the programs used, the sequence datasets and the quality of the alignments, leading to difficulties in comparing independent studies performed with different tools. Phylogenetic analyses performed on different fragments are useful to inform the identification of breakpoints, whose positions are supported by the demonstration of different clustering of fragments located apart each breakpoint. This is the strategy we employed.

Other studies have characterized the mosaicism of sarbecoviruses through recombination analysis. For example the program 3SEQ[1] was used to study the interlineage recombinations of SARS-CoV-2 variants[2]. 3SEQ allowed also to show that Sarbecovirus evolution was shaped by multiple recombination events, including several at the beginning and within the spike, as we observed in our study[3]. 3SEQ allowed the identification of hotspots of recombination in sarbecoviruses, most of them located at the beginning of the spike[4]. On the other hand, GARD appeared more conservative than RDP5,which includes a large set of algorithms, to identify breakpoints[5]. They identified 21 breakpoints through the genome of SARS-CoV-2, among which some are also found in our analysis (for example the one at the beginning of the spike, which seems to correspond to our breakpoint no 9). These

comments along with corresponding references have been added as comments of the Supplementary Figure S2.

*1. Lam, H. M., Ratmann, O. & Boni, M. F. Improved Algorithmic Complexity for the 3SEQ Recombination Detection Algorithm. Mol. Biol. Evol. **35**, 247–251 (2018).*

*2. Jackson, B. et al. Generation and transmission of interlineage recombinants in the SARS-CoV-2 pandemic. Cell **184**, 5179-5188.e8 (2021).*

*3. Boni, M. F. et al. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Nat. Microbiol. **5**, 1408–1417 (2020).*

*4. Forni, D., Cagliani, R. & Sironi, M. Recombination and Positive Selection Differentially Shaped the Diversity of Betacoronavirus Subgenera. Viruses **12**, E1313 (2020).*

5. Lytras, S. et al. Exploring the natural origins of SARS-CoV-2 in the light of recombination. 2021.01.22.427830 https://www.biorxiv.org/content/10.1101/2021.01.22.427830v3 (2021) doi:10.1101/2021.01.22.427830.

*6. Lemoine, F. et al. NGPhylogeny.fr: new generation phylogenetic services for non-specialists. Nucleic Acids Res. **47**, W260–W265 (2019).*

4. There is scant information on the methods for the phylogenetic analysis. What model of nucleotide substitution was used? Why weren't bootstrap values inferred for such a small dataset.

The substitution model considered for all phylogenetic reconstructions was GTR+G. PhyML was used to infer phylogenies at the whole genome and at the breakpoints nucleotide levels. For RBD phylogeny presented in Figure 1, LG+G model was considered. The tree topology search was performed by selecting the best of NNI (Nearest Neighbor Interchange) and SPR (Subtree Pruning and Regraphing). Instead of performing classical bootstrapping, the statistical test selected for branch support was the approximate Bayes parameter, as proposed in the NGPhylogeny portal[6]. Only posterior probabilities >0.5 were considered.

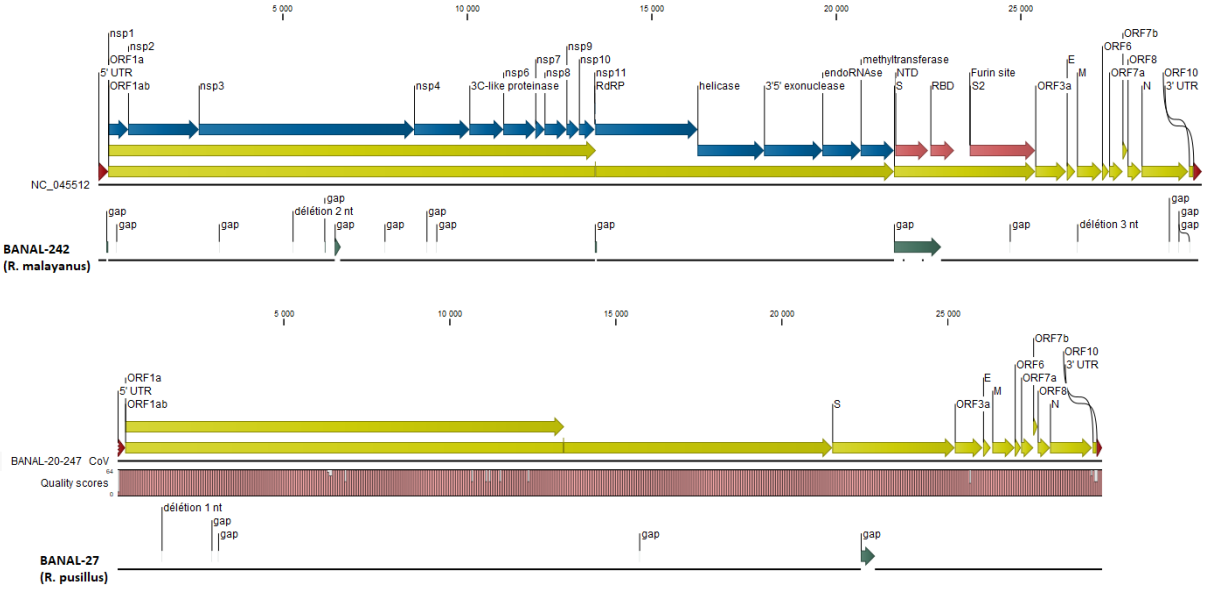
5. Can you describe the coverage you achieved in sequencing, particularly for key sites of interest?

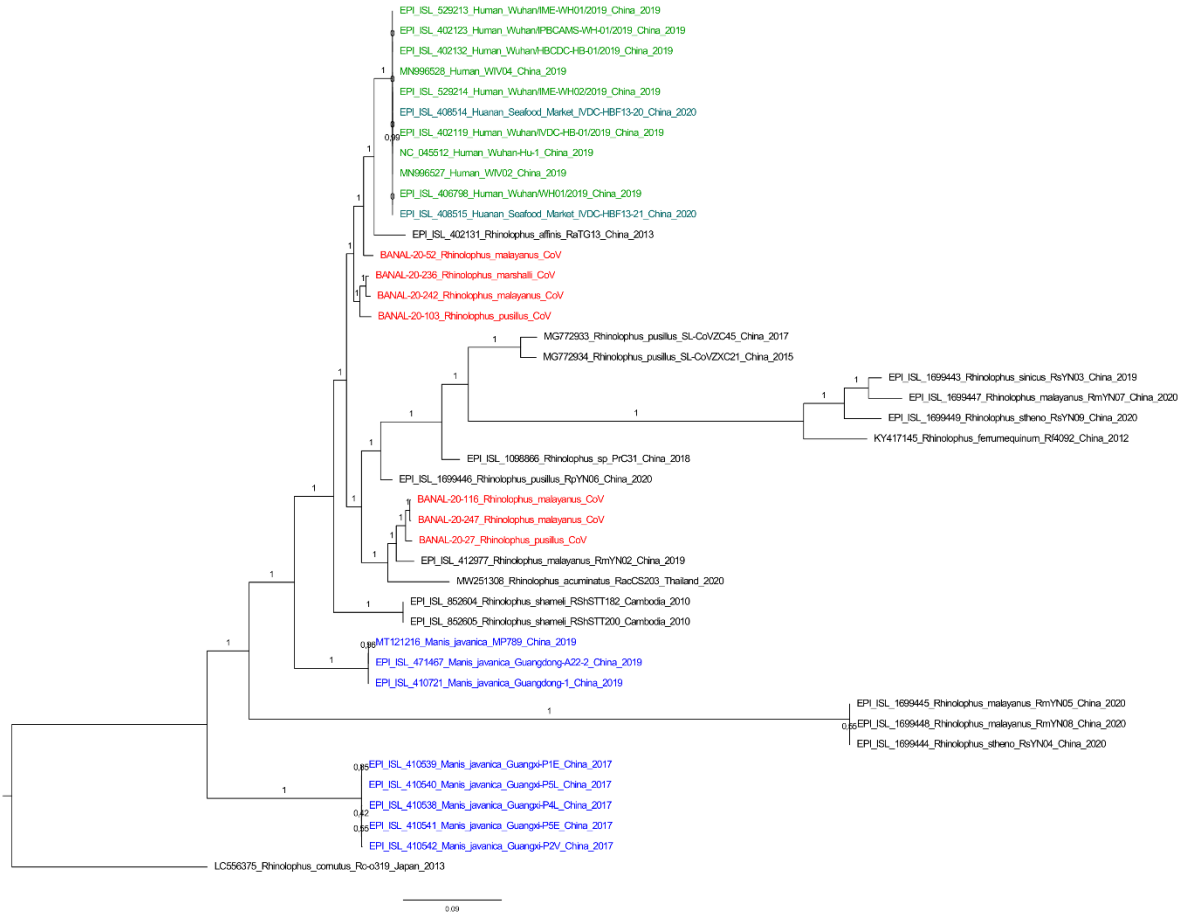
*Below is presented the average vertical coverage obtained **after NGS** for all BANAL samples. Of note, several key regions (noted by a * in the Table) were obtained by Sanger sequencing after multiple overlapping PCRs, resulting in a lower vertical coverage of the region but with higher confidence of the consensus sequence. In addition, BANAL-52 was sequenced in three independent NGS runs, including one AmpliSeq, to achieve a good genome coverage. Therefore, the coverage presented for BANAL-52 results from the combination of the three runs. This information was added to the Supplementary data, in Table S8.*

BANAL sample	Average coverage for whole genome (SD)	Average coverage in Spike (SD)	Average coverage in RBD (SD)
BANAL-52	83720 X (89558) *	75247 X (97210) *	36623 X (64721) *
BANAL-103	43 X (47) *	19 X (16) *	0.42 X (0.74) *
BANAL-236	2872 X (2477)	4100 X (2862)	2847 X (1530)
BANAL-116	662 X (512)	1038 X (511)	679 X (309)
BANAL-247	15 X (13) *	19 X (14) *	7 X (3) *

6. Lines 95-97: it seems like a waste to throw out data from viruses with 90% genome coverage, given how few data points we have. Was there good coverage of the spike protein?

We totally agree that the finishing of the two remaining bat sarbecoviruses is important to increase the knowledge of bat SARS-CoV-2-related coronaviruses. Unfortunately (probably due to low viral loads in the rectal swabs) we were unable to obtain the complete genome of these viruses, especially regarding the S1 domain of spike (see below the remaining gaps) despite several NGS runs and Sanger sequencings. We decided not to include these viruses in the final analysis because of the lack of sequencing of key residues in the spike; but partial phylogenetic analyses performed at the whole genome level showed that BANAL-242 (from Rhinolophus malayanus) is close to BANAL-103 and BANAL-236 while BANAL-27 (from Rhinolophus pusillus) is close to BANAL-116 and BANAL-247 (see below).





7. Line 189-190: how conserved were SARS-CoV-2 epitopes in the BANAL viruses?

There are several epitopes targeted by several antibodies. If we just consider the RBD which is the target of most of the neutralizing antibodies, the most characterized ones are those competing for the ACE2 binding site and that region is very conserved with only 1 or 2 mismatches. We are currently testing a set of different Mabs. We cannot predict the effect of the two mismatches in their binding affinity but there should be many anti SARS-CoV-2 antibodies binding BANAL RBDs. The rest of the RBD, out of the RBM region, should not have major changes in the epitopes because it is well conserved.

Manuscript organization:

1. A summary table is needed in the main text that shows how the BANAL viruses share important residues in the RBD that contact with ACE2, compared to other bat and pangolin viruses, rather than burying this key data in the supplement (show data for cleavage site as well).

As suggested, we updated Figure 1 and replaced the whole genome phylogeny by the phylogenetic tree of the RBD region of representative sarbecoviruses. Unfortunately, we did not find enough space to include a readable figure with contact residues as a part of the Figure 1 panel. The information is however available in Supp. Figure 3.

2. It would also be helpful to have some simple summary table or chart of genetic similarity between Laos viruses, SARS-CoV-2, RaTG13 and select other bat viruses, including (a) whole-genome, (b) spike, (c) RBD, etc.

The Supp. Figure 1, which presented the nucleotide and amino-acid identity matrices of representative sarbecoviruses, was updated to add the matrix at the whole genome level and to divide the Spike gene into NTD, RBD and S2 domains as suggested.

3. The Discussion section is unnecessarily long and at times hyperbolic and needs to be tightened and focused. For example, this study provides some important insights but it is inaccurate to suggest in line 221 that these data show that RaTG13 is no longer the "proximal ancestor" of SARS-CoV-2. For quite some time it's been shown that RaTG13 only contributed sections of the SARS-CoV-2 genome and other viruses like RmYN02 are most closely related in other sections. Also rephrase line 247. The BANAL viruses are similar to SARS-CoV-2 in an important way but they are not "very close".

We have reduced the length of the discussion and eliminated some redundancies within the paragraphs of this section and with the introduction section. We have also deleted the sentence "and that R. affinis RaTG13 is no longer considered the proximal ancestor of SARS-CoV-2". Regarding the mention "very close", according to ICTV criteria "Viruses that share more than 90% aa sequence identity in the conserved replicase domains are considered to belong to the same species. This 90% identity threshold serves as the sole species demarcation criterion". Amino acid identity to SARS-CoV-2 of the replicase of BANAL 52-236-103 is superior to 99.1-99.9 % (99.9% for BANAL-52). Also, BANAL-52 whole genome is 96.8% identical at the nucleotide level to SARS-CoV-2 and shares a similar RBD. We think therefore that the term "very close" reflects the reality without being emphatic and we'd prefer to keep it.

<u>Minor clarifications:</u>

1. It would be helpful in the supplement to have maps of the geographic ranges of these Rhinolophus bat species (mashalli, malaynus, and pusillus)

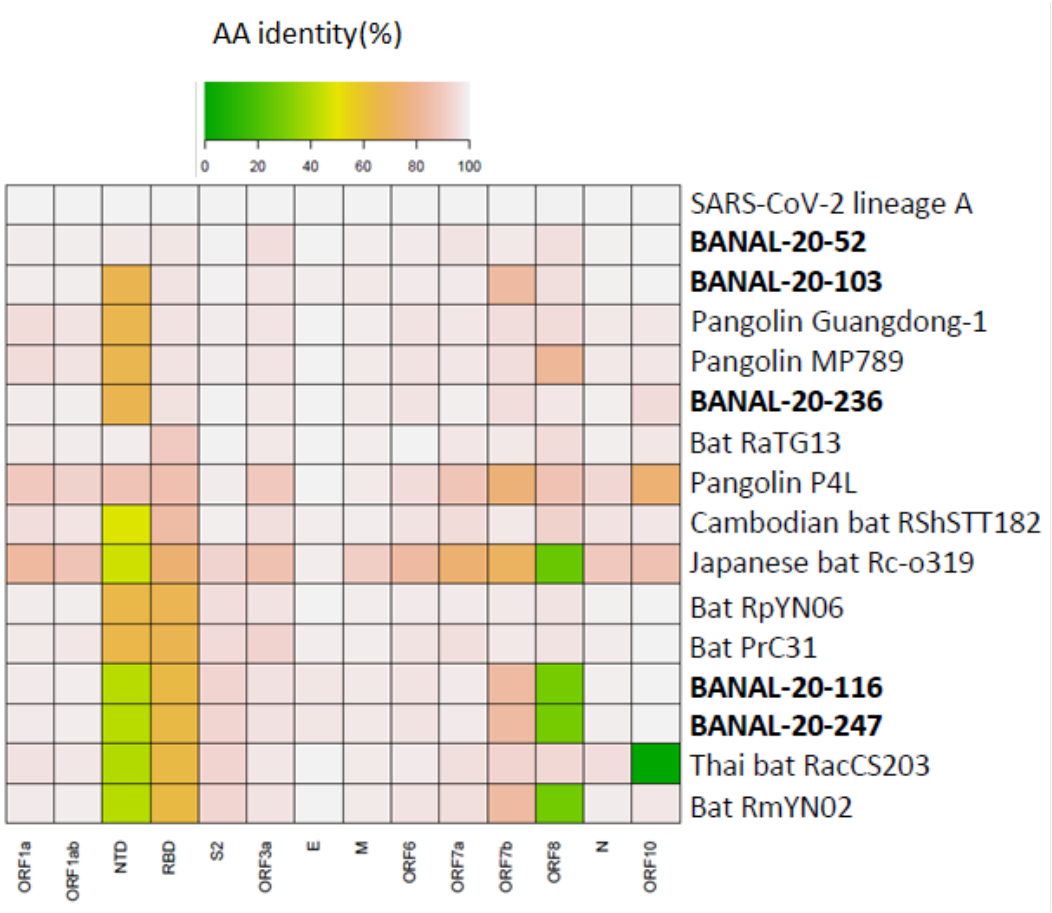
We have followed this suggestion and the map is available as Supp. Fig 11.

2. There is a typo in the Figure S4 legend: "clivage site" should be "cleavage site"

The legend was corrected accordingly.

3. The resolution in Figure 1D is poor. There should be more resolution for the higher part of the scale. Otherwise all that is being highlighted is the lower AA identity data points, which are not of interest.

Despite numerous attempts (one attempt is presented below) we were unable to derive a picture with such a variable scale. Tabular values are available in the supplementary section for the readers who need a higher level of precision (Figure S1).



4. This line in the Abstract is not accurate: “Several studies have suggested the involvement of pangolin coronaviruses in SARS-CoV-2 emergence7–9 Better to simply say that viruses with similar genomic features as SARS-CoV-2 have been identified in pangolins.

The abstract has been modified accordingly.

5. Did you also test the bat saliva or urine samples?

Since the submission, we have tested 307 available saliva samples (PCR only) and we plan to test urine samples. We have found 6 positive saliva samples including 2 already positive on fecal swabs. Saliva from bat-52, 103, 236, 116 and 247 were negative. We have not modified the text as the results are not comprehensive and do not modify the main message.

6. Line 57: “evolved only through sporadic mutations” not true, recombination (Jackson et al 2021 Cell)

This has been corrected and the reference has been cited.

7. GenBank accession numbers are listed for the 5 BANAL viruses that are the focus of this study. Did you sequence and deposit the genomes of the other 19 coronaviruses from bats? These are valuable data. The next bat coronavirus that infects humans could come from any part of the tree and it’s helpful to have as many background sequences as possible as context.

We agree that any bat coronavirus could emerge in humans, and therefore all genetic data of animal coronavirus is of great interest for the prediction of spillover events. We are currently finishing the genomes of the 6 decacoviruses, 5 pedacoviruses, 1 rhinacovirus and 6 nobecoviruses. We are also trying to fill the remaining gaps of the 2 sarbecoviruses. The genomes will be deposited into public repositories (GenBank and GISAID) as soon as possible.

Referee #5 (Remarks to the Author):

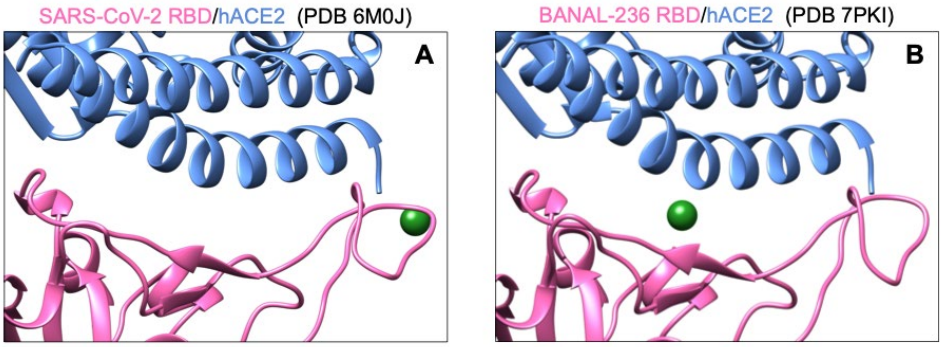
The work reported in this manuscript sought to identify bat viruses similar to SARS-CoV-2 in order to further probe the origin of the virus that led to the COVID-19 pandemic. I have been asked to provide peer-review on its computational component, which focused on three such identified viruses (BANAL 52, 103, 236).

Based on the X-ray structure 6M0J, which describes the SARS-CoV-2 spike RBD complexed with a fragment of the ACE2 receptor, homology models were constructed for BANAL 52/103 and 236, and atomistic simulations were performed in triplicate for the models and 6M0J. Given sequence identity ~97% within the modeled regions, homology modeling is highly appropriate. Reported MODELLER and GROMACS protocols are reasonable, and sampling time reaches the microsecond timescale.

We thank the reviewer for the appreciation of the computational part of this manuscript.

It was unclear to me why the authors retained crystallographic ions in the models, but not waters, which can be important for mediating interactions in complexes.

We thank the reviewer for this observation. Zinc is essential to the catalytic activity of the zinc metallopeptidase angiotensin-converting enzyme 2 (ACE2). A zinc ion was indeed resolved in the crystal structure used as template of our homology models of BANAL RBDs/hACE2 complexes (PDB 6M0J). This ion is coordinating residues H374, H378, and E402. For this reason, we decided to retain the Zinc atoms when building our homology models. Regarding waters, we agree with the reviewer that in general these molecules can be important in mediating protein-protein interactions. However, in the crystal structure of the SARS-CoV-2 RBD/hACE2 complex only one water molecule was resolved close to the RBD/hACE2 interface (see figure below, panel A, water oxygen as green sphere). As we could not predict how the mutations in the BANAL sequences would affect the interface with hACE2, we preferred to not include this crystallographic water molecule in the construction of the homology models. We verified a posteriori if water molecules were present at the interface of BANAL-236 RBD/hACE2 in our crystal structure (PDB 7PKI), and indeed we found one water molecule, but in a different position with respect to 6M0J (see figure below, panel B, water oxygen as green sphere).



It was unclear why they did not perform their simulations at physiological temperature.

The aim of our MD simulations is to provide an atomistic description of the interactions at the interface of the BANAL-RBDs with hACE2 that could explain the strong binding affinities measured in the BLI experiments. For this reason, we decided to perform our simulations at room temperature as the BLI experiments were carried out at 298K. We have added a sentence in the Supplementary Methods to explain the reason for our choice.

A more significant concern is that all of the models describe fragments, which are at high risk of unfolding in the absence of restraints, especially on long timescales, especially for a system that contains unstructured regions like the RBD; yet, the authors do not indicate the use of restraints during their production runs.

The RBDs (~200 residues) and hACE2 constructs (~600 residues) used in our simulations were the same used in the BLI experiments. Ad hoc conformational restraints to prevent unfolding of the two subunits would have introduced artifacts in our MD simulations and affected the comparison with experiments. Furthermore, on the basis that:

- *SARS-CoV-2, BANAL-52/103, and BANAL-236 RBDs all formed stable complexes with hACE2 with binding affinities in the nanomolar range,*
- *the structure of the SARS-CoV-2 RBD / hACE2 complex was determined at 2.45 Å resolution, with the majority of residues resolved except for a handful of flexible residues in the N-terminal and C-terminal region of the RBD,*

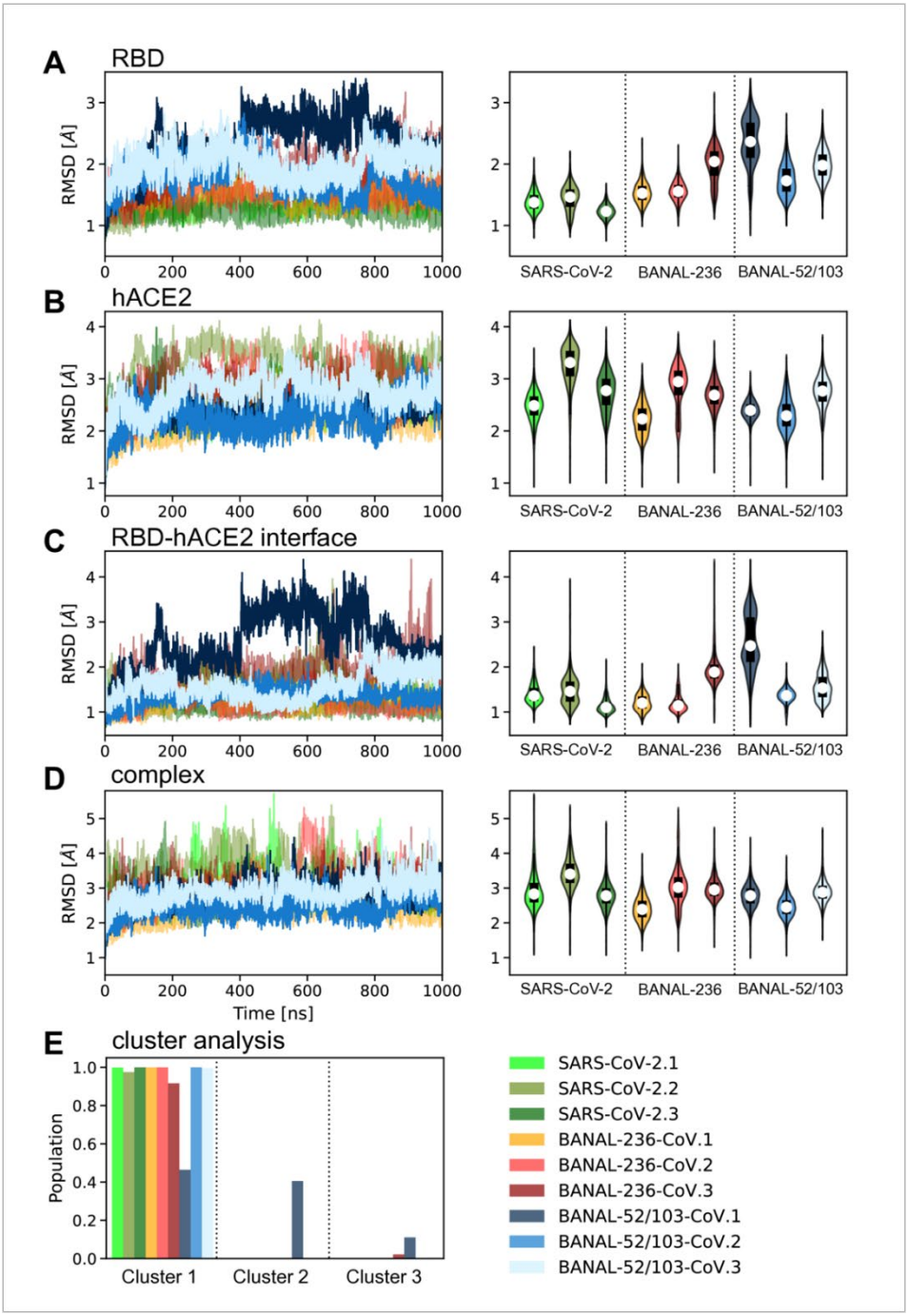
We were confident that the complexes studied in our MD simulations would have been stable in the microsecond time scale. In summary, the addition of conformational restraints to our MD simulations would have compromised the comparison with experimental data and was not necessary to ensure the stability of the complex.

They do describe the use of distance restraints “to ensure the formation of 7 disulfide bonds” in the template, and this is very odd. A disulfide bond should be an actual S-S bond, with specific force field parameters to define its bond-angle-dihedral behavior; disulfide bonds are indeed available in CHARMM36m. While these distance-based restraints likely helped to prevent unfolding of the RBD fragment, the disulfide dynamics are modeled incorrectly if treated as restraints, and with so many for so small a fragment this could influence the dynamics of the entire RBD-ACE2 model.

There must be a misunderstanding here, possibly due to our description of the methods not being sufficiently clear. During the course of our MD simulations, disulfide bonds were correctly treated in the framework of the CHARMM36m force field, as mentioned by the reviewer. Disulfide bond restraints were only applied by MODELLER via the Model.path_ss_templates() patch (see https://salilab.org/modeller/manual/node189.html). 7 disulfide bonds were detected in the template 6M0J (3 in hACE2 and 4 in the RBD) and enforced in the generation of the homology models by CHARMM-like distance and dihedral angles restraints. The PDB files of the resulting models contained 7 SSBOND entries that were subsequently read by CHARMM-GUI so that the correct CHARMM36m force field parameters for the linked cysteines were used. We thank the reviewer for this observation that prompted us to clarify the text in the revised version of the Supplementary Methods.

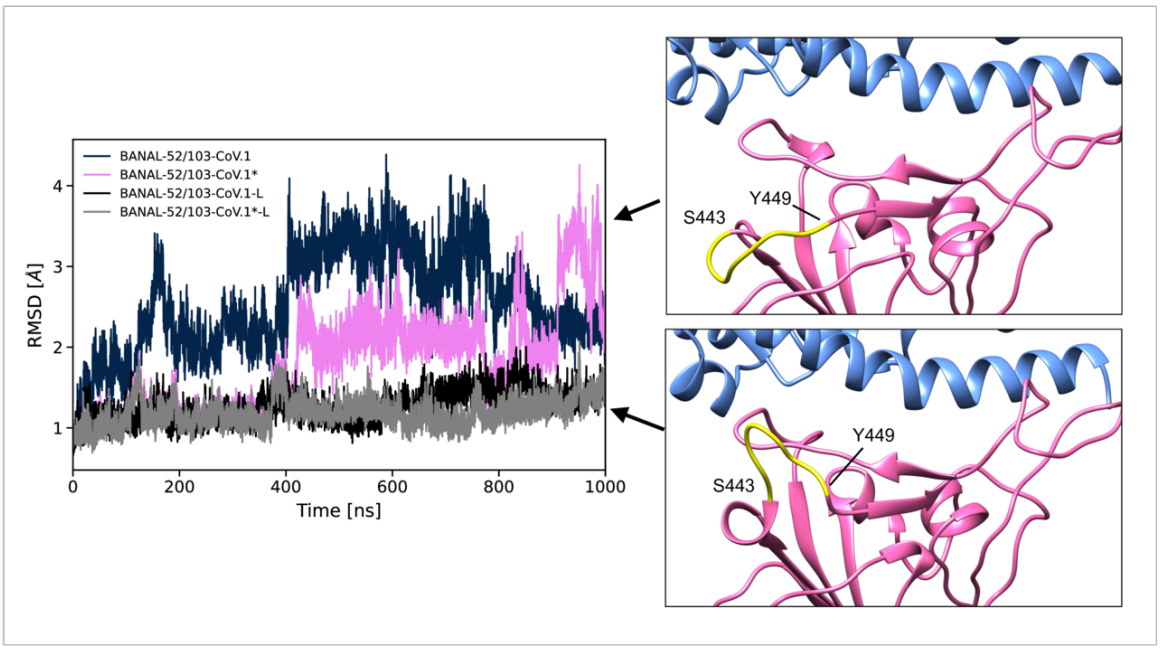
The reader could be assured that the simulations are stable and the production runs converged if plots of RMSD over time were presented, particularly for the entire complex. The current RMSD plots are presented as distributions, and not all of the distributions are visible in the plots presented due to overlap. One cannot know from this presentation if the RMSD is converging, diverging, or simply fluctuating over the course of the simulation.

Time series of RMSD from the starting conformation are in general not a good measure of the convergence of MD simulations. They can instead provide indications of the stability of the conformation used as starting point of the simulation, if this is what the reviewer meant with this comment. In this regard, we agree with the reviewer that an overlay of the RMSD distributions is probably not the most effective visualization. We therefore modified all the RMSD plots in Figure S5 to include a time series of the RMSD for all the simulations performed (first column of panels A, B, C, and D) along with violin plots to visualize the RMSD distributions on the ensembles determined in our simulations (second column of panels A, B, C, and D). We believe that these violin plots are a clearer representation that facilitates the comparison between different RMSD distributions compared to our original presentation. We have also calculated the backbone RMSD for the entire complex (panel D), as requested by the reviewer. We did not report this quantity in the original Figure S5 because we believe that the RMSD on the entire complex is not very a sensitive probe of conformational changes at the interface of the two subunits (compare for example the time series of the RMSD for the BANAL-52/103-CoV.1 simulation in panel C and D). The modified Figure S5 is reported below.



Still it is clear from the irregular and broad RMSD distributions that the BANAL 52/103 interface has quite different behavior from the other models. In the absence of further information, this reader questions the stability of the interface. The different behavior of BANAL 52/103 leads to the identification of two other clusters that describe the interface, and the reason for this is not discussed–this is a key point, if the one model exhibits multiple clusters, its behavior is not in fact "identical" to the others, as the authors claim.

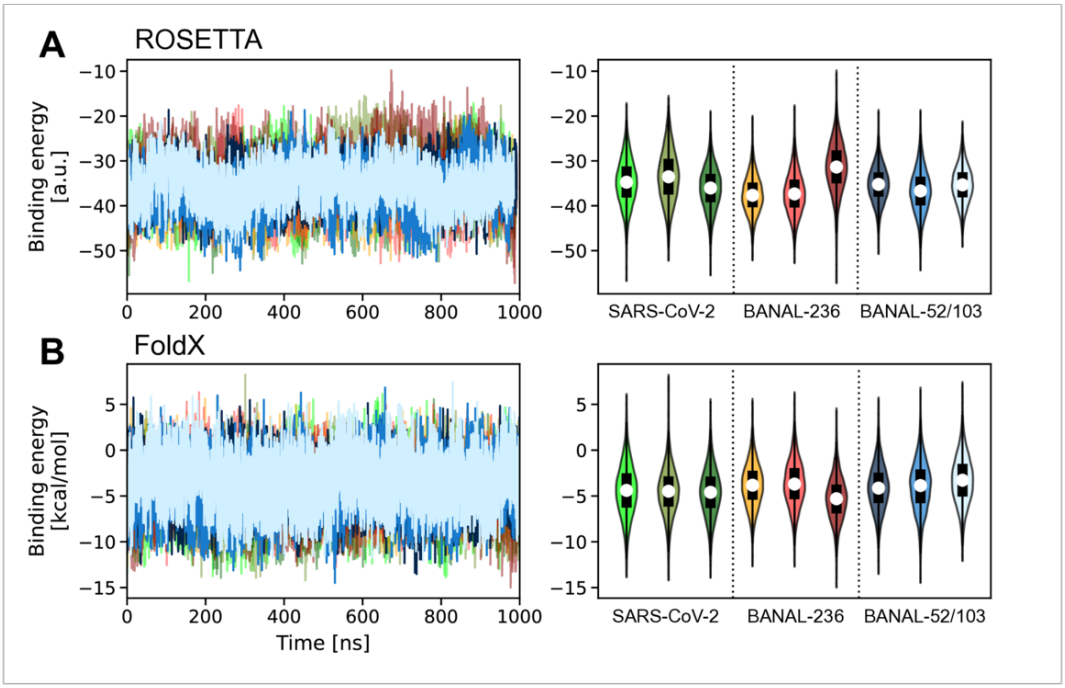
We do not agree with the reviewer's statement that "the BANAL 52/103 interface has quite different behavior from the other models". In fact, only one out of 3 simulations of the BANAL-52/103 RBD/hACE2 complex (marked as BANAL-52/103-CoV.1 in the previous figure) showed large fluctuations of the RBD-hACE2 interface RMSD. Notably, the time series of the RMSD for the entire



Another way to communicate simulation findings to readers would be through the presentation of a supplemental movie of the simulation, which would be welcome here. A number of other MD simulations have already been published from 6M0J; perhaps the authors could support their case also by citing others.

We thank the reviewer for this suggestion. Movies are certainly nice and appealing visualizations of MD trajectories. However, we believe that our new analysis in terms of time series of RMSD and ROSETTA/FoldX scores suggested by the reviewer and their corresponding violin plots along with our original analysis of salt bridges and hydrogen bonds formation frequency provide: i) more informative representations of the structural and dynamical properties of the systems studied; ii) a more effective way to compare the behavior of different RBD/hACE2 constructs.

The authors use ROSETTA and FoldX scores to approximate binding energies and report these data as distributions, which are again difficult to decipher because of overlap. For what is visible, there appears to be a range of values across which the distributions are centered. An alternative presentation would allow readers to judge for themselves how similar the data really are.



After carrying out all of this computational work, the authors make very little comment on it, note no caveats or limitations, and do not refer to it during discussion.

This manuscript presents the results of a large variety of experimental and computational approaches, including next-generation sequencing, biochemistry and in cell measurements, and experimental structural biology studies, and the implications of our findings on the origin of SARS-CoV-2. We believe that our MD simulations provide important insights for the comprehension of the high affinity for hACE2 of the receptor binding domains found in the coronaviruses isolated in bats of North Laos. We extensively commented on the results of our simulations and compared them with our BLI binding assays as well as crystallographic data in the structural biology (results) section of the paper. We have now added a brief sentence referring to the molecular simulations during discussion. Unfortunately, there is not much space left here to further expand our discussion on the structural biology aspects of our work, and we will consider a separate publication on this topic.

While the computational aspect of this work is not novel, and does not contribute significantly to the conclusions of the study, it appears sufficiently reliable for its purpose. While the actual interaction between the SARS-CoV-2 spike and ACE2 receptor depends on many other factors that are not considered in the simplified computational models used here, the work does indicate that the isolated binding interfaces of SARS-CoV-2 and BANAL 52/103 and 236 share some notable similarities over one-microsecond timescales.

We thank the reviewer again for the appreciation of the computational part of this work and we hope that our improved presentation of the results will help clarify their doubts.

Reviewer Reports on the First Revision:

Referee #1:

The authors have satisfactorily addressed all points raised by the reviewers.

Referee #3:

he authors have adequately addressed the points raised by this reviewer. The impact of the study has been much increased by the new finding that the virus replicates in the human cell lines Caco-2 and Calu-3.

One minor point should still be addressed by modifying the text: Formally, blockade of entry by soluble ACE2 is not proof that the virus uses ACE2 for entry, please see PMID: 34214467.

Referee #4:

The authors have addressed all of my previous comments.

Referee #5:

This is an important study that establishes that bat coronaviruses exist in nature that bind efficiently to human ACE2 receptors, resolving a key question in the evolutionary origins of SARS-CoV-2. The revised manuscript is clearer and improved. The only remaining issue is the incorrect alignment in the region around the cleavage site.

It's fine to acknowledge some uncertainty in the alignment in the region around the S1/S2 cleavage site. But defaulting to the MAFFT alignment often leads to the wrong alignment. That is the point of Spyros's virological.org post the authors refer to. Spyros presents an alignment for the BANAL viruses identical to what I was suggesting (see Figure 2 in his more recent "One Year Later"

addition from Nov 24 2021), with an insertion in the cleavage site for BANAL-20-116 and BANAL-20-247, similar to RmYN02 (also see alignment from Zhou et al 2020). The alignment in Figure S4 should be corrected to resemble the one in Spyros's post and to be consistent with previous publications, including the Holmes review in Cell. The text should be updated accordingly to mention this insertion in the cleavage site, similar to what was seen in RmYN02. I don't understand the authors' resistance, since this observation enhances the paper and presents another similarity between the bat viruses and SARS-CoV-2.

Minor:

1. The authors fixed the line in the original abstract "Several studies have suggested the involvement of pangolin coronaviruses in SARS-CoV-2 emergence" but lines 132-133 in the introduction still read: "Several studies also suggested the role of pangolin-related coronaviruses in the emergence of SARS-CoV-2".

2. There are some errors in the text for Figure S2.
"datatsets"
"the position of a given sequence differ"
"at the breakpoints nucleotide levels"

3. Line 57: "recombinations" is odd phrasing. Better to say "sporadic mutations and recombination" or "sporadic mutations and recombination events"

4. Line 153: entry "into" human cells

5. Nature titles are limited to 75 characters. Maybe just use "Discovery of bat coronaviruses in Indochinese peninsula that bind to hACE2"

Referee #6:

With respect to reviewer comments on the computational work presented in the manuscript, the authors have been quite thorough in addressing critiques. Indeed, most of my concerns arose from misunderstanding the authors' description of their methods, but now they have made edits and additions that provide clarity on these points. With these concerns addressed, I find the computational work was executed with care and attention to detail. I also find the updated data visualizations to be much more clear and informative; the presented findings sufficiently support the authors' conclusions. Altogether, this reviewer is satisfied with the revisions and looks forward to seeing the authors' work published.

Author Rebuttals to First Revison:

Referees' comments:

- Referee #1 (Remarks to the Author):

The authors have satisfactorily addressed all points raised by the reviewers.

- Referee #3 (Remarks to the Author):

The authors have adequately addressed the points raised by this reviewer. The impact of the study has been much increased by the new finding that the virus replicates in the human cell lines Caco-2 and Calu-3.
One minor point should still be addressed by modifying the text: Formally, blockade of entry by soluble ACE2 is not proof that the virus uses ACE2 for entry, please see PMID: 34214467.

We thank the reviewer for this information and agree that alternative receptor(s) may play a role in BANAL-236 infection of Vero cells. Of note, 1/ virus entry was strongly inhibited by soluble ACE2, supporting a major role of ACE2 in virus binding to Vero cells; 2/ BANAL-52, -103 and -236 viruses present a E and not a D in position 484, and Puray-Chavez et al. reported that the E484D mutation is mandatory for H522 infection through an ACE2-independant pathway.
Since alternative routes of entry may still exist, we added a sentence in the manuscript as the reviewer suggested (lines 264-265).

- Referee #4 (Remarks to the Author):

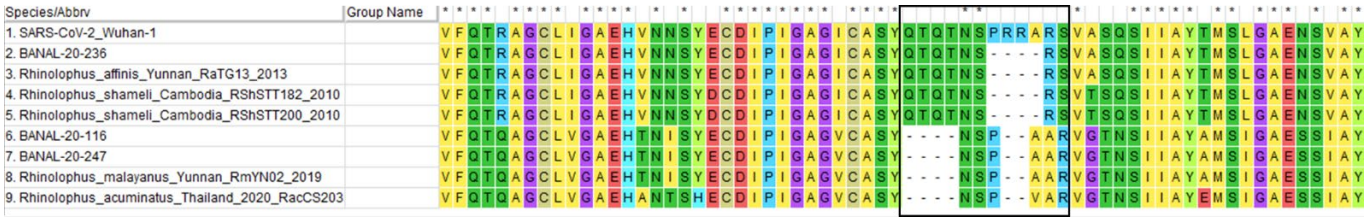
The authors have addressed all of my previous comments.

- Referee #5 (Remarks to the Author):

This is an important study that establishes that bat coronaviruses exist in nature that bind efficiently to human ACE2 receptors, resolving a key question in the evolutionary origins of SARS-CoV-2. The revised manuscript is clearer and improved. The only remaining issue is the incorrect alignment in the region around the cleavage site.
It's fine to acknowledge some uncertainty in the alignment in the region around the S1/S2 cleavage site. But defaulting to the MAFFT alignment often leads to the wrong alignment. That is the point of Spyros's virological.org post the authors refer to. Spyros presents an alignment for the BANAL viruses identical to what I was suggesting (see Figure 2 in his more recent "One Year Later" addition from Nov 24 2021), with an insertion in the cleavage site for BANAL-20-116 and BANAL-20-247, similar to RmYN02 (also see alignment from Zhou et al 2020). The alignment in Figure S4 should be corrected to resemble the one in Spyros's post and to be consistent with previous publications, including the Holmes review in Cell. The text should be updated accordingly to mention this insertion in the cleavage site, similar to what was seen in RmYN02. I don't understand the authors' resistance, since this observation enhances the paper and presents another similarity between the bat viruses and SARS-CoV-2.

We agree with the reviewer that identifying a bat coronavirus carrying sequences that could act as precursors of a furin cleavage site would greatly improve the understanding of the origin of SARS-CoV-2. BANAL-52, -103, and -236 viruses, which are to date the closest viruses to SARS-CoV-2 in the RBD region, do not present this type of sequence. BANAL-116 and -247 viruses are phylogenetically more distant and belong to a clade that include Chinese RmYN02 and Thai RacCS203 viruses.
Regarding the alignment around the S1/S2 cleavage site with representative bat SARS-CoV-2-like viruses, we think that both alignments are possible (the one presented in the current version of our article that correspond to the alignment of nucleotides at the codon level, and the one proposed by the reviewer). We tried to reproduce the alignment proposed by the reviewer with different alignment methods (including MAFFT with three different iterative refinement methods [auto = L-INS-I, E-INS-I, G-INS-i]; and MUSCLE algorithms), but all alignment attempts result in the alignment presented in our

article. Only ClustalW aligner (gap open penalty 10; gap extension penalty 0.1 or 0.2 for pairwise or multiple comparison respectively) was able to reproduce partially the reviewer alignment (see below). In addition, Spyros Lytras presented in his post that his alignment was manually corrected (virological.org/t/the-sarbecovirus-origin-of-sars-cov-2-s-furin-cleavage-site/536/6).



Therefore, we decided to present both alignments, each at the nucleotide and at the amino-acid levels, (lines 254-259, extended data 4) for a comprehensive information of the readers.

Minor:

1. The authors fixed the line in the original abstract “Several studies have suggested the involvement of pangolin coronaviruses in SARS-CoV-2 emergence" but lines 132-133 in the introduction still read: “Several studies also suggested the role of pangolin-related coronaviruses in the emergence of SARS-CoV-2“.

We apologized for this mistake and modified the introduction accordingly (lines59-60).

2. There are some errors in the text for Figure S2.
“datatsets”
“the position of a given sequence differ”
“at the breakpoints nucleotide levels”

Corrections were made accordingly.

3. Line 57: “recombinations” is odd phrasing. Better to say “sporadic mutations and recombination” or “sporadic mutations and recombination events”

The sentence was rephrased accordingly.

4. Line 153: entry “into” human cells

Correction was made accordingly.

5. Nature titles are limited to 75 characters. Maybe just use “Discovery of bat coronaviruses in Indochinese peninsula that bind to hACE2”

We changed the titer to “Discovery of bat coronaviruses close to SARS-CoV-2 and infectious for human cells” It is 81 characters long. We can delete “and” and it will become “ Discovery of bat coronaviruses close to SARS-CoV-2 infectious for human cells” so 77 characters long. We hope that one of the two titles could be acceptable. If not “bat coronaviruses close to SARS-CoV-2 infectious for human cells” (64 characters) would be fine

- Referee #6 (Remarks to the Author):

With respect to reviewer comments on the computational work presented in the manuscript, the authors have been quite thorough in addressing critiques. Indeed, most of my concerns arose from misunderstanding the authors' description of their methods, but now they have made edits and additions that provide clarity on these points. With these concerns addressed, I find the computational work was executed with care and attention to detail. I also find the updated data visualizations to be much more clear and informative; the presented findings sufficiently support the authors' conclusions. Altogether, this reviewer is satisfied with the revisions and looks forward to seeing the authors' work published.